Toxic Speech Detection in Portuguese: A Comparative Study of Large Language Models

Amanda S. Oliveira and **João P. R. Alvarenga**
Graduate Program in Computer Science
Federal University of Ouro Preto
35.400-000 – Ouro Preto – MG – Brazil
amanda.oliveira2,joao.alvarenga@aluno.ufop.edu.br

Thiago C. Cecote and **Vander L. S. Freitas** and **Eduardo J. S. Luz**
Department of Computing
Federal University of Ouro Preto
35.400-000 – Ouro Preto – MG – Brazil
thiago.cecote@aluno.ufop.edu.br
vander.freitas,eduluz@ufop.edu.br

Abstract

This research addresses the automatic detection of toxic speech in Portuguese. Utilizing the ToLD-Br dataset, which includes 21,000 annotated tweets, we examine the performance of Large Language Models (LLMs) such as OpenAI's ChatGPT and the monolingual MariTalk from Maritaca AI. The study focuses on their effectiveness in identifying Toxic speech, the influence of few-shot learning, and the intricacies of annotating datasets, particularly regarding vulgar language (swear words). Our experiments reveal that MariTalk (Sabiá) demonstrates a nuanced understanding of colloquial Portuguese. Meanwhile, ChatGPT, especially when augmented with few-shot learning, shows robustness comparable to baseline methods. This investigation underscores the value of both monolingual and lower-capacity models in the nuanced field of language-specific Toxic speech detection, offering insights into their competitive edge against models like ChatGPT.

1 Introduction

In 2023, X (formerly Twitter) updated its documentation on hateful conduct (Twitter, 2023), clearly defining what they consider a violation of this policy. This includes explicit prohibitions against messages that promote fear and discrimination against specific groups. Additionally, the policy considers the repeated use of insults, degrading stereotypes, or images that dehumanize a particular group as violations. In light of these updated policies, developing effective automatic hate and toxic speech detection strategies becomes increasingly crucial.

Automated toxic speech detection strategies typically involve linguistic feature analysis, lexicon-based approaches, and supervised machine learning algorithms trained on labeled datasets (Schmidt and Wiegand, 2017; Vargas et al., 2022b). Advanced techniques, including natural language processing and deep learning methods, seek to comprehend the semantics and context of textual content (Leite et al., 2020; Vargas et al., 2022a). Yet, substantial challenges persist due to the complexity of human language, the fast evolution of toxic speech, and the balance needed between free speech and the fight against harmful content.

Moreover, while research has predominantly focused on English, there has been notable progress in detecting toxic speech in Portuguese. For instance, the ToLD-Br dataset (Leite et al., 2020), containing 21,000 annotated tweets, allows for new advancements. Despite BERT-based models reaching macro-F1 scores between 70% and 80% on this dataset, room for improvement exists.

The use of Large Language Models (LLMs) has gained significant notoriety due to the success of OpenAI's ChatGPT. Today, impressive results are being achieved using LLMs for various natural language tasks (Kocoń et al., 2023), including for Portuguese, such as answering questions from the Brazilian National High School Exam (Silveira and Mauá, 2018; Nunes et al., 2023), text reading and comprehension (FaQuAD) (Sayama et al., 2019), and social network sentiment analysis (Brum and Nunes, 2017), prediction of depressive disorder (dos Santos and Paraboni, 2023), among others. A comprehensive study by Kocoń et al. (2023) demonstrated how ChatGPT, via OpenAI's API, can be competitive for various NLP tasks, including hate speech. In Oliveira et al.

(2023), authors showed the efficacy of ChatGPT-3.5 Turbo, using a zero-shot approach, for detecting toxic speech in Portuguese. The same study indicated that other supervised learning methods struggle with test data from different distributions, whereas ChatGPT is more resilient in this regard. However, OpenAI's ChatGPT is a model with a large number of parameters and, consequently, high computational cost. This study centers its investigation on the analysis of toxicity and hate speech in Portuguese texts. Thus, this work aims to explore smaller, Portuguese-specialized language models, such as Sabiá from Maritaca AI [1]. Sabiá is a monolingual language model trained for Portuguese (Pires et al., 2023) and available via a free API (MariTalk API) as a chatbot. Unlike Oliveira et al. (2023), we also investigate the few-shot approach for ChatGPT 3.5 and the Maritalk here. Additionally, we deeply analyze the ToLD-Br dataset, considering the annotation challenges discussed in Poletto et al. (2021), focusing on texts containing vulgar language. In this work, we concentrate on three research questions:

- Q1: How does the performance of a monolingual Large Language Model (LLM) for Portuguese (MariTalk-Sabiá) compare to a multilingual counterpart (ChatGPT) in the detection of toxic speech?
- Q2: What is the efficacy of a few-shot learning approach in enhancing the performance of LLMs for hate/toxic speech detection?
- Q3: What are the challenges associated with dataset annotation for toxic speech detection, and how does including vulgar and obscene language (swear words) affect the performance of these models?

Through four experiments, the study analyzed models' proficiency in processing Portuguese for toxic text detection. MariTalk-Sabiá demonstrated notable efficacy, especially when enhanced by the few-shot approach, and showed a more sophisticated understanding of colloquial Portuguese. Even with lower capacity, monolingual models can be a promising way to solve the problem addressed here.

[1]API MariTalk: https://www.maritaca.ai/

2 Detection of Hate Speech and Toxicity in Portuguese

The effective detection of hate speech and toxicity in Portuguese texts presents unique challenges due to the diverse speakers across various countries. While each region and social group exhibits distinct cultural differences, they contribute to the complexity of hate speech detection in Portuguese. Comprehensive and representative datasets are essential to address this challenge effectively. However, there is a relative scarcity of labeled data in Portuguese compared to English, which significantly impedes the development of robust detection systems. In this context, analyzing existing datasets becomes critical to identifying representative content that captures the multifaceted nature of hate speech in Portuguese across diverse cultural and regional contexts. The lack of a common taxonomy connecting various concepts related to toxic or hateful speech also poses a challenge, leading to possible biases and misclassification issues in detection models (Poletto et al., 2021). Below, we highlight four datasets and works of interest.

2.1 OffcomBr

The dataset proposed in de Pelle and Moreira (2017) collects comments from a news site (G1[2]). A total of 1,250 comments were manually annotated by three different annotators, using the Fless Kappa measure to gauge the level of agreement among them.

The authors provided two different sets, named OFFCOMBR-2 and OFFCOMBR-3. The difference between them is that OFFCOMBR-2 includes comments considered offensive by at least two annotators, while OFFCOMBR-3 consists of comments on which all three annotators agreed. Besides, the dataset was also classified among racism, sexism, homophobia, xenophobia, religious intolerance, and insults. The most frequent class is "insults". The authors established a baseline using n-grams and infoGain as features and used Naive Bayes and Support Vector Machine (SVM) classifiers. The SVM-based models performed better than others, achieving a weighted F-score in the range of 77-82.

2.2 HLPHSD

The HLPHSD dataset, detailed in Fortuna et al. (2019), is a corpus of 5,668 tweets from 1,156 users

[2]https://g1.globo.com/

collected between January and March 2017. Annotation started with non-specialist volunteers who categorized tweets as hate or non-hate, followed by experts assigning nuanced labels to create an 81-category hierarchical taxonomy. Cohen's Kappa coefficient ensured consistency among annotators. The dataset's inclusion of Brazilian and European users captures the nuances of the Portuguese language, with 31.5% of tweets classified under hate speech.

The authors of the dataset employed pre-trained embeddings and Long Short-Term Memory (LSTM) networks to establish a baseline. This evaluation resulted in an F1-score of 78%.

2.3 Hate-Br

The Hate-Br database, introduced in Vargas et al. (2022a), comprises 7,000 Instagram comments in Brazilian Portuguese, annotated by three expert annotators. The annotation was structured in three layers: binary (offensive vs. non-offensive), level of offensiveness (highly, moderately, and slightly offensive), and specific hate speech categories (xenophobia, racism, homophobia, sexism, religious intolerance, partisanship, apology to the dictatorship, anti-Semitism, and fatphobia).

For baseline establishment in Vargas et al. (2022a), the authors utilized n-grams and bag-of-n-grams with TFIDF preprocessing for data representation, applying Naive Bayes, SVM, Multilayer Perceptron, and Logistic Regression for classification. The dataset was split into 80% for training, 10% for testing, and 10% for validation. The study achieved an F-score of 85% in hate speech detection and 78% in offensive speech detection.

2.4 ToLD-Br

The ToLD-Br dataset, introduced in Leite et al. (2020), serves as a specialized corpus for detecting toxic language within Brazilian Portuguese on Twitter/X. This dataset was collected over the months of July and August 2019, employing a dual-strategy approach to maximize the inclusion of potentially toxic content. The first strategy targeted tweets containing predefined terms associated with toxicity, while the second strategy broadened the scope by capturing tweets directed at influential figures, likely to attract abusive responses. The resultant dataset is comprehensive, encompassing 21,000 tweets that were anonymized and then rigorously annotated by three independent volunteers to ensure a diverse and representative compilation of various forms of toxic language, including LGBTphobia, racism, misogyny, and xenophobia. The final corpus, with 60% of posts derived from keyword-focused strategies and the remainder from threads involving public figures, was partitioned with an 80% allocation for training and a stratified 20% reserved for testing.

Research conducted in Leite et al. (2020) showcased the efficacy of BERT-based models on this dataset, yielding a macro-F1 score of 76% in hate speech detection. This underscores the significance of expansive monolingual datasets in enhancing computational model precision. Meanwhile, the study in da Rocha Junqueira et al. (2023) revealed the superiority of BERTimbau for toxic speech detection within ToLD-Br, substantiating the selection of BERTimbau as the baseline model for the present research.

Further exploration in Oliveira et al. (2023) focused on ChatGPT in zero-shot mode for detecting toxic speech in the ToLD-Br test partition. Though ChatGPT 3.5 Turbo did not surpass established baseline models, it showed comparable outcomes. The study also highlighted the significant effect of data distribution variations in baseline methods. While ChatGPT demonstrated resilience to these variations, it has constraints related to high financial costs, limited accessibility, and undisclosed details about the Reinforcement Learning from Human Feedback (RLHF) phase. This paves the way for a shift in focus toward investigating an open architecture model, such as the Llama-based model tailored for Portuguese with 65 billion parameters called Sabiá (Pires et al., 2023). Although an instance of it cannot be accessed directly, it can be accessed via a free API.

3 Experimental Methodology

This study employs the GPT 3.5 models from OpenAI via a paid API[3] and also utilizes the MariTalk from Maritaca AI, accessible through a free API[4]. The models evaluated include gpt-3.5-turbo-0613 and an instance of Sabiá-65B Architecture. A zero-temperature setting is used for all language models, meaning more deterministic inferences. As a baseline, a BERT-based model trained for Portuguese is used.

[3]https://platform.openai.com/

[4]https://github.com/maritaca-ai/maritalk-api

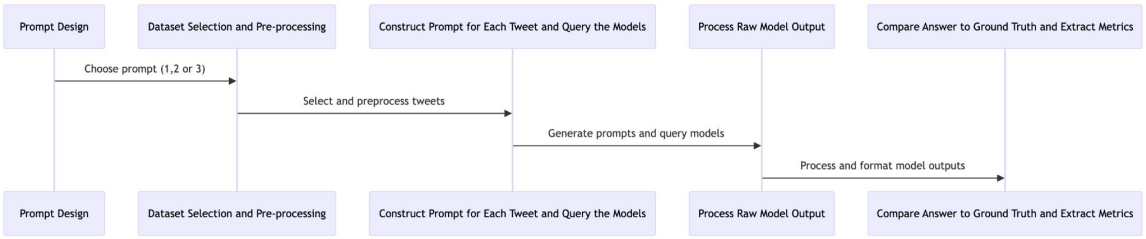

Figure 1: Methodological flow for evaluating LLMs APIs.

3.1 Evaluation with LLMs via APIs

In evaluating prompts with LLM chatbots using APIs, two models, ChatGPT and MariTalk, were assessed according to the methodology outlined in Figure 1.

ChatGPT, a significant progression in the GPT series based on the Transformer architecture, showcased notable advancements from its predecessor models. The initial GPT version utilized the Transformer decoder stack with unidirectional attention, expanding its capabilities to tasks like translation, summarization, and question answering (Radford et al., 2018). GPT-2 further extended these functions by doubling the input context length, increasing parameters, and enhancing training data volume for better task-specific learning. The subsequent model, GPT-3, with its 175 billion parameters and training on vast textual data, excelled in zero-shot and few-shot scenarios, demonstrating substantial improvements (Brown et al., 2020). The most recent innovation, InstructGPT, refined the model to better cope with human needs, leveraging Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022). This evolution reflects the ongoing commitment to align language models, such as ChatGPT, with real user needs (Radford et al., 2018; Ouyang et al., 2022).

In contrast, MariTalk employs the same architecture as Sabiá-65B; although trained on an undisclosed set of data, the model capacity is known. The instance of Sabiá-65B presented in Pires et al. (2023), based on the Llama 65B language model, was evaluated on several tasks from the Poeta benchmark, including text classification, gap filling, and translation. Sabiá outperformed English-centric and multilingual language models in many of these tasks, setting new benchmarks for performance in Portuguese natural language processing tasks. Although Pires et al. (2023) provides details about the LLM, it is still not open to the public. Also, other processes behind the MariTalk API, such as the use of reinforcement learning from human feedback or fine-tuning processes for instruction, remain undisclosed.

3.1.1 Prompt Design

The prompts used in this study were adapted from the work of Oliveira et al. (2023) for comparative purposes. In our approach, two types of prompts were explored for few-shot and zero-shot, and a third prompt, proposed in this work, was employed for a zero-shot analysis. Prompts 1 and 2 were used to assess the ability of MariTalk and ChatGPT 3.5 Turbo models to identify toxic texts. It is noteworthy that while ChatGPT 3.5 Turbo uses the concept of a "system" message, in the case of MariTalk, this message was directly incorporated into the prompt. Prompt 3 was used exclusively with MariTalk.

The prompts, along with the dataset instances sourced from Portuguese-speaking users, were translated and thoughtfully adapted into English for the manuscript readers, but the models received input texts in Portuguese.

Prompt #1

For the *zero-shot* method, the prompt is structured as follows:

> "You are conducting text analysis. The following text was used in a social media context. Tell me, with a yes or no, if you consider this text toxic: [text]."

For the *few-shot* method, with n instances per class, the dialogue is structured as follows:

> User: "You are conducting text analysis. The following text was used in a social media context. Respond only with 'yes' or 'no' if the text below contains hate speech, offense, aggression, or other forms of toxicity: [example text]."
>
> Assistant: "Yes, it is toxic." (or "No, it is not toxic.")

The "example text" here means an input from the training set.

Prompt #2

For the *zero-shot* method, the prompt is structured as follows:

> "You are conducting text analysis. The following text was used in a social media context. Respond only with 'yes' or 'no' if the text below contains hate speech, offense, aggression, or other forms of toxicity: [text]"

For the *few-shot* method, with n instances per class, the dialogue is structured as follows:

> User: "You are conducting text analysis. The following text was used in a social media context. Respond only with 'yes' or 'no' if the text below contains hate speech, offense, aggression, or other forms of toxicity: [example text]."
>
> Assistant: "Yes, it is toxic." (or "No, it is not toxic.")

The "example text" here means an input from the training set.

Prompt #3

Deriving from the prompts presented in Oliveira et al. (2023), we introduce here a version of the prompt with explicit commands, instructing the model to classify it as toxic if the text already contains obscene and aggressive texts. The prompt is structured as follows:

> "You are conducting text analysis. The following text was used in a social media context. Respond only with 'yes' or 'no' if the text below contains hate speech, offense, aggression, or other forms of toxicity. If the text includes swear words, low-level language, or obscene words, classify it as offensive, therefore toxic: [text]."

3.1.2 Dataset Selection and Preprocessing

In this stage, a dataset comprising a series of tweets is chosen for evaluation (ToLD-Br). Each tweet underwent individual processing and anonymization; specifically, mentions of users were uniformly replaced with "@user". Furthermore, all embedded links within these tweets were systematically removed.

3.1.3 Constructing the Prompt and Querying the Models

During this phase, each tweet is carefully tailored to fit the designated prompt (specified in the [text] field) before initiating a query to either the MariTalk or ChatGPT API. To efficiently manage potential network disruptions, a strategy of retries with exponential backoff was implemented.

3.1.4 Processing the Raw Response

Responses to each tweet were diligently recorded and subsequently refined to align with the output format of the baseline model for comparison. In this process, responses explicitly stating "yes" were converted to a numerical value of 1.0, while those indicating "No" were translated to 0.0. It's noteworthy that responses from large language models can sometimes deviate from expected patterns, as demonstrated in the following examples:

> Query: "You are conducting text analysis. The following text was used in a social media context. Respond only with 'yes' or 'no', if you consider this text toxic: ridiculous the attitude of both, mercy on such people"
>
> *ChatGPT answer: "No."*
>
> Query: "You are conducting text analysis. The following text was used in a social media context. Respond only with 'yes' or 'no', if you consider this text toxic: rt @user wrote, didn't read, the stick hit."
>
> *ChatGPT answer: "Yes, it is toxic."*

Thus, regular expressions (regex) are used here to check if a given text contains the words "yes" or "no" as whole and independent words within the sentences.

3.1.5 Metrics

The evaluation encompasses a range of metrics, notably the class-specific F1-score, precision, and recall, as well as their macro and weighted counterparts. The macro variant of these metrics computes the metric separately for each class before averaging them, thereby ensuring equal representation for all classes. Conversely, the weighted variant also calculates these metrics individually for each class but applies a weighting in the averaging process proportional to the class's prevalence. These metrics are defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{2}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{3}$$

$$\text{Macro F1-score} = \frac{1}{N} \sum_{i=1}^{N} \text{F1-score}_i, \tag{4}$$

$$\text{Weighted F1-score} = \sum_{i=1}^{N} w_i \times \text{F1-score}_i, \tag{5}$$

where TP, FP, and FN correspond to true positives, false positives, and false negatives, respectively; N is the number of classes, and w_i is the proportion of the total sample size that class i represents.

3.2 Baseline Methods

For the baseline model, we followed the methodology proposed in Leite et al. (2020) and Oliveira et al. (2023), also employing a BERT-based model. We used the *simpletransformers* library [5] with default arguments for reproducibility. The pre-trained model was BERTimbau (Souza et al., 2020)[6].

3.3 Methodology for Analysis and Verification of Dataset Annotations

To better understand the nuances of our results, we conducted a detailed review of the dataset annotations (test set). We explored the hypothesis that the presence of swear words might influence text classification as hate speech or toxic discourse, which could reflect common challenges in annotation consistency.

To conduct this investigation, we first compiled a list encompassing various categories of swear words, totaling 60 terms and expressions, including spelling errors and internet neologisms. Subsequently, we used this list to identify sentences containing such terms in the test data, resulting in 1,010 instances in the test data. These instances were then re-annotated by a specialist, who followed a specific guide covering all contexts of swear word usage in Brazilian Portuguese. It is worth noting that, in the Portuguese language, words commonly considered obscene can function as adjectives, interjections, or intensity adverbs, as exemplified in (Original source follow in italics for further reference):

- Fucking delicious cake. *"Bolo gostoso pra caralho."* (intensity adverb)
- Blessed be the mute, damn it! *"Bendito seja o mute, caralho!"* (interjection)
- Bahia is so fucking awesome, giants. *"O bahia é foda demais pqp, gigantes."*(adjective/interjection)

To refine our understanding of the dataset's nuances, we conducted a thorough review of the annotations within the test set. During this process, we identified 380 instances where the presence of swear words, often used as interjections or intensifiers, may have led to their initial categorization as toxic content. We carefully reassessed these instances. After a considered re-evaluation, we updated the labels where necessary, resulting in a revised test set that we believe reflects a more colloquial interpretation of the language used.

[5]https://simpletransformers.ai/

[6]https://huggingface.co/neuralmind/bert-large-portuguese-cased

4 Results and Discussion

For the experiments involving ChatGPT and MariTalk, the official APIs were utilized. In total, 24,984 prompts were sent to the MariTalk API and 16,656 to the OpenAI API. Post-processing of each prompt's response was conducted following the query, with necessary adjustments made for comparison against the baseline model, BERTimbau. Regarding the baseline, the BERTimbau model was locally trained on a machine equipped with an NVIDIA 3090 GPU with 24GB, an Intel(R) Core(TM) i9-10900 CPU @ 2.80GHz, and 128GB of RAM. Four experiments were conducted to address the three research questions posed in this study. The source code for reproducing the experiments is available at `https://github.com/ufopcsilab/ToxicSpeech-Propor2024`.

4.1 EXP 1: Assessing the Impact of a Portuguese-Specific Language Model: MariTalk

To evaluate a monolingual model's performance for the task, we compared the results obtained from the MariTalk API against the top outcomes reported in Oliveira et al. (2023). For this purpose, both the experiments with BERTimbau and the ChatGPT 3.5 Turbo API were re-implemented and tested under the same setup. The radar chart of Figure 2 compares the performance of three models, with axes representing precision, recall, and F1 scores for toxic and non-toxic categories. The chart indicates that the MariTalk model is particularly precise at identifying non-toxic texts, meaning it has a lower rate of falsely labeling non-toxic texts as toxic. However, its ability to recognize toxic texts (recall for toxic) and its overall accuracy and balance between precision and recall (F1 scores for both toxic and non-toxic) might not be as strong as some of the other models represented on the chart.

4.2 EXP 2: Examining the Impact of the Few-Shot Approach

Experiment 2 aimed to address research question Q2, specifically investigating the impact of employing a few-shot approach on the models under study. For this experiment, instances from the training dataset were randomly selected from both classes in a balanced manner and used to compose

Model	Prompt	Precision	F-score
BERTimbau	—-	0.76	0.75
ChatGPT 3.5-turbo zeroshot	prompt 2	0.74	0.74
ChatGPT 3.5-turbo + 10 fewshots	prompt 1	0.74	0.56
ChatGPT 3.5-turbo + 10 fewshots	prompt 2	0.75	0.72
Maritaca zeroshot	prompt 1	0.70	0.50
Maritaca zeroshot	prompt 2	0.73	0.69
Maritaca + 10 fewshots	prompt 2	0.73	0.73
Maritaca + 20 fewshots	prompt 2	0.74	0.72
Bertimbau#	—-	0.72	0.73
ChatGPT 3.5-turbo zeroshot#	prompt 2	0.72	0.72
Maritaca zeroshot#	prompt 1	0.70	0.55
Maritaca zeroshot#	prompt 2	0.68	0.67

Table 1: Summary of results on ToLD-Br test set. Methods marked with # were evaluated on a re-annotated ToLD-Br test set.

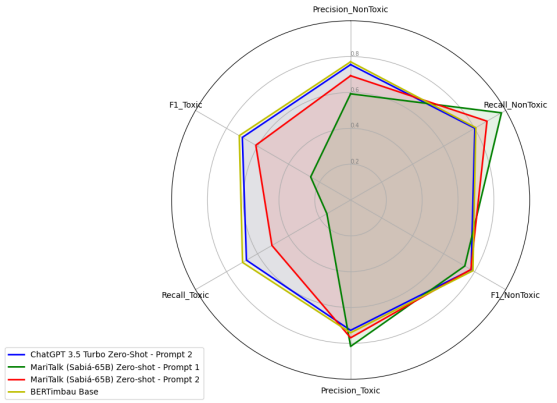

Figure 2: Zero-shot experiments for Q1.

the prompts. We experimented with 10 and 20 instances per class. Figure 3 displays a comparison of the best results achieved, and Table 1 shows a more complete panorama. It's important to note that BERTimbau is included for comparison purposes, as it remained unchanged between tests.

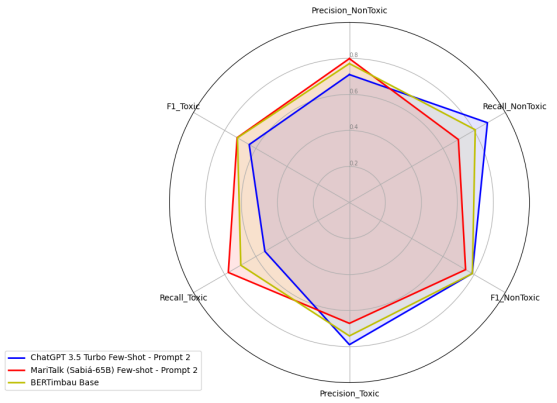

Figure 3: Few-shot experiments for Q2.

In the zero-shot modeling context, we observed that certain popular words and expressions, often categorized as slang or vulgar language, were not automatically classified as toxic, aggressive, or hateful language. The following phrases, extracted from the ToLD-Br test partition, were initially classified as non-toxic by MariTalk in zero-shot mode but were reclassified when the few-shot approach was employed:

- "get out of this, they've already done a lot but now some super badass girls are coming and they're playing like hell."
- "but in the end I understood all the shit and I got along with the guys... I loved it, everyone was fucking awesome yesterday."
- "I dreamed that I was dating?? Fuck, the guy was hot for me."
- "mami calling me a bitch lol lol so good."
- "bro, I'm fucking mad at this network!!!!"

In our opinion, these phrases highlight MariTalk-zero-shot's ability to discern between colloquial language and potentially offensive language, demonstrating an advanced understanding of the usage of words and slang in different contexts.

With the few-shot approach, MariTalk began to classify these instances correctly, or rather, align them with the dataset ToLD-Br's labels.

4.3 EXP 3: Investigating a Third Prompt: Focus on Aggressive and Obscene Words

Two experiments were conducted to address research question Q3, experiments 3 and 4. Experiment 3 aimed to understand the impact of incorporating specific commands into the prompt to force the classification of instances as toxic whenever an aggressive or obscene word appeared. Figure 4 shows the effect of prompt 3 on MariTalk's classification.

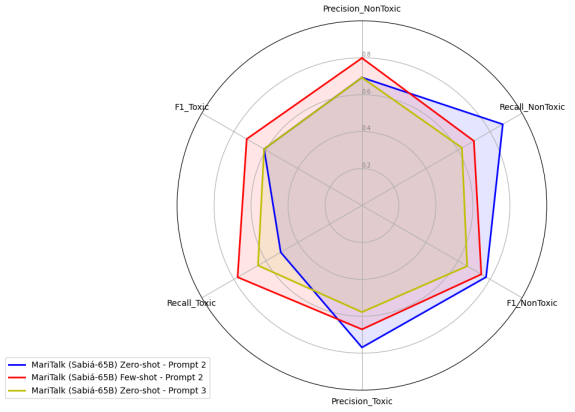

Figure 4: Few-shot experiments for Q3.

4.4 EXP 4: Analyzing Annotations from the ToLD-Br Test Partition

To perform a comparative performance analysis between the models, taking into account the reannotation of the test data according to our methodology, we proceeded with experiments using the new test set. The results demonstrated a small discrepancy, as can be seen in Table 1. In this case, we are interested in the response of the models without any interference, that is, zero-shot.

MariTalk, while achieving satisfactory outcomes in the initial experiment, experienced a notable decline in F-score and precision when the reannotated test set was applied with prompt 2. We believe that prompt 2 instructs the model to become more sensitive. ChatGPT also presents a decline in terms of F-score and precision for prompt 2. We would like to highlight that the MariTalk model improved precision for the case of the first prompt.

4.5 Discussion

The four experiments conducted provided valuable insights into the performance and utility of the models in processing Portuguese text. Experiment 1, focusing on the monolingual MariTalk chatbot, demonstrated its effectiveness in handling Portuguese language tasks, as evidenced by its comparison with top results from previous studies. The introduction of the few-shot approach in Experiment 2 marked a significant improvement in MariTalk's ability to correctly classify instances, particularly those involving colloquial and slang expressions, highlighting the model's improved understanding and contextual interpretation with additional examples.

Experiments 3 and 4 further explored the subtleties of language model performance. Experiment 3 examined the impact of a prompt specifically designed to identify aggressive and obscene words, showing the model's sensitivity to prompt design and its influence on classification accuracy. In Experiment 4, the analysis of reannotated test data from the ToLD-Br dataset indicated a slight discrepancy in the performance of ChatGPT and MariTalk. MariTalk exhibited increased precision, supporting the hypothesis that it better understands the nuances of colloquial Portuguese.

5 Conclusion

The experiments conducted provided insights into the performance and adaptability of both ChatGPT and MariTalk in processing Portuguese for toxic text detection. Both models demonstrated competitive performances, yet neither managed to outperform the BERTimbau model when applied to the ToLD-Br dataset. Notably, MariTalk, being a monolingual model with an open Llama architecture, showed particular promise. The study also revealed that employing a few-shot approach, even with as few as ten example instances per class, significantly influenced the results. However, it is crucial to recognize the limitations of our study, particularly the lack of in-depth access to the models, which might have impacted our findings. Moving forward, a valuable path for research could involve direct interaction with Large Language Models (LLMs), bypassing the constraints of API-based access.

Acknowledgments

We would like to express our sincere thanks to the company Blip, whose generous support and invaluable assistance were crucial for the presence of two authors at this event. The authors would also like to thank the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* - Brazil (CAPES) - Finance Code 001, *Fundacão de Amparo à Pesquisa do Estado de Minas Gerais* (FAPEMIG, grants APQ-01518-21), *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq, grants 308400/2022-4), and Universidade Federal de Ouro Preto (PROPPI/UFOP) for supporting the development of this study.

References

T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry,

A. Askell, et al. 2020. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pages 1877–1901.

Henrico Bertini Brum and Maria das Graças Volpe Nunes. 2017. Building a sentiment corpus of tweets in brazilian portuguese. *arXiv preprint arXiv:1712.08917*.

Júlia da Rocha Junqueira, Claudio Luis Junior, Félix Leonel V Silva, Ulisses Brisolara Côrrea, and Larissa A de Freitas. 2023. Albertina in action: An investigation of its abilities in aspect extraction, hate speech detection, irony detection, and question-answering. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 146–155. SBC.

Rafael P de Pelle and Viviane P Moreira. 2017. Offensive comments in the brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. SBC.

Wesley Ramos dos Santos and Ivandré Paraboni. 2023. Predição de transtorno depressivo em redes sociais: Bert supervisionado ou chatgpt zero-shot? In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 11–21. SBC.

Paula Fortuna, João Ricardo da Silva, Leo Wanner, and Samuel Nunes. 2019. A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online*, pages 94–104.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*, page 101861.

João Antônio Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924.

Desnes Nunes, Ricardo Primi, Ramon Pires, Roberto Lotufo, and Rodrigo Nogueira. 2023. Evaluating gpt-3.5 and gpt-4 models on brazilian university admission exams. *arXiv preprint arXiv:2303.17003*.

Amanda S Oliveira, Thiago C Cecote, Pedro HL Silva, Jadson C Gertrudes, Vander LS Freitas, and Eduardo JS Luz. 2023. How good is chatgpt for detecting hate speech in portuguese? In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 94–103. SBC.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. In *Intelligent Systems*, pages 226–240, Cham. Springer Nature Switzerland.

Fabrício Poletto, Valerio Basile, and Manuela Sanguinetti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources & Evaluation*, 55(2):477–523.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Hélio Fonseca Sayama, Anderson Viçoso Araujo, and Eraldo Rezende Fernandes. 2019. Faquad: Reading comprehension dataset in the domain of brazilian higher education. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 443–448. IEEE.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10. Association for Computational Linguistics.

Igor Cataneo Silveira and Denis Deratani Mauá. 2018. Advances in automatically solving the enem. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 43–48. IEEE.

Fabricio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*, pages 403–417. Springer.

Twitter. 2023. Hateful conduct policy. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

Felipe Vargas, Isabela Carvalho, Felipe R de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022a. Hatebr: A large expert annotated corpus of brazilian instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183.

Francielle Vargas, Fabiana Rodrigues de Góes, Isabelle Carvalho, Fabrício Benevenuto, and Thiago Alexandre Salgueiro Pardo. 2022b. Contextual-lexicon approach for abusive language detection.